AFRL-RH-WP-TR-2015-0007

FOREIGN LANGUAGE ANALYSIS AND RECOGNITION (FLARE) PROGRESS

Brian M. Ore
Stephen A. Thorn
David M. Hoeferlin
SRA International
5000 Springfield Street, Suite 200
Dayton, OH, 45431

Raymond E. Slyh
Eric G. Hansen
Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Human Trust and Interaction Branch
2255 H Street
Wright-Patterson AFB, OH, 45433

February 2015
Interim Report for 1 October 2012 to 30 November 2014

Distribution A: Approved for public realease: distribution is unlimited.

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING
HUMAN EFFECTIVENESS DIRECTORATE
HUMAN-CENTERED ISR DIVISION
HUMAN TRUST AND INTERACTION BRANCH
WRIGHT-PATTERSON AFB OH 45433
AIR FORCE MATERIAL COMMAND
UNITES STATES AIR FORCE

STINFO COPY

NOTICE AND SIGNATURE PAGE

AFRL-RH-WP-TR-2015-0007 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

// signature //
Raymond E. Slyh
Work Unit Manager
Human Trust and Interaction Branch

// signature //
Louise A. Carter, Ph.D.
Human-Centered ISR Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)* 1–02–2015	**2. REPORT TYPE** Interim	**3. DATES COVERED** *(From — To)* 1 October 2012 – 30 November 2014

4. TITLE AND SUBTITLE

Foreign Language Analysis and Recognition (FLARe) Progress

5a. CONTRACT NUMBER
FA8650–09–D–6939

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)

[1]Brian
[1]Steve Thorn
[1] Dave Hoeferlin
[2]Raymond E. Slyh
[2]Eric G. Hansen

5d. PROJECT NUMBER
5328

5e. TASK NUMBER
0028

5f. WORK UNIT NUMBER
H06K (5328X02S)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

[1]SRA International
5000 Springfield Street, Suite 200
Dayton, OH 45431

[2]Air Force Research Laboratory
7llth Human Performance Wing
Human Effectiveness Directorate
Human Trust and Interaction Branch
Wright-Patterson AFB, OH 45433

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Materiel Command
Air Force Research Laboratory
711 Human Performance Wing
Human Effectiveness Directorate
Human Centered ISR Division
Wright-Patterson AFB OH 45433

10. SPONSOR/MONITOR'S ACRONYM(S)
711 HPW/RHXS

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RH-WP-TR-2015-0007

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

88ABW-2015-2172; Cleared 29 April 2015

14. ABSTRACT

This interim report provides research results in the areas of automatic speech recognition (ASR) and information retrieval (IR).

15. SUBJECT TERMS

Automatic speech recognition (ASR), information retrieval (IR).

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Raymond E. Slyh
U	U	U	SAR	37	**19b. TELEPHONE NUMBER** *(include area code)*

THIS PAGE INTENTIONALLY LEFT BLANK.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SUMMARY

This document provides a summary of work completed by government researchers and SRA International under the work unit H06K (5328X02S), Foreign Language Analysis and Recognition (FLARe). This work was performed over the period 1 October 2012 to 30 November 2014 under contract FA8650-09-D-6939.

The following tasks were completed on Automatic Speech Recognition (ASR). Korean language models (LMs) were developed to reduce the number of Out-of-Vocabulary (OOV) words encountered by the recognizer. Levantine Arabic and Farsi ASR systems were trained on conversational telephone speech. Three different methods were investigated for combining Pashto ASR systems. Software was developed for training and evaluating hybrid deep neural network (DNN) hidden Markov model (HMM) speech recognition systems. An English ASR system was developed for the International Workshop on Spoken Language Translation (IWSLT) 2013 evaluation. Six different techniques were investigated for interpolating LM probabilities. Finally, English and Italian ASR systems were developed for the IWSLT 2014 evaluation.

Improvements were made to the Haystack Multilingual Multimedia Information Extraction and Retrieval (MMIER) system that was initially developed under a prior work unit. Major additions to the user interface include the following: support for uploading multiple files, expansive changes to the media player, additional Machine Translation (MT) capabilities, and integration of geolocation information. Scripts were developed for translating web pages and displaying the results in the same format as the input. Research into HTML5 was initiated to improve functionality across different operating systems. The processing pipeline was updated to provide support for decoding hybrid DNN-HMM systems, support for N-gram and Recurrent Neural Network (RNN) LM rescoring, and improved text extraction from Portable Document Format (PDF) files. Japanese, Chinese, and Pashto speech recognition systems were developed and then incorporated into Haystack.

1.0 INTRODUCTION

This document provides a summary of work completed by government researchers and SRA International under the work unit 5328X02S, Foreign Language Analysis and Recognition (FLARe). This work was performed over the period 1 October 2012 to 30 November 2014 under contract FA8650-09-D-6939.

The following tasks were completed on automatic speech recognition (ASR). Korean language models (LMs) were developed to reduce the number of out-of-vocabulary (OOV) words encountered by the recognizer. Levantine Arabic and Farsi ASR systems were trained on conversational telephone speech. Three different methods were investigated for combining Pashto ASR systems. Software was developed for training and evaluating hybrid deep neural network (DNN) hidden Markov model (HMM) speech recognition systems. An English ASR system was developed fprp the International Workshop on Spoken Language Translation (IWSLT) 2013 evaluation. Six different techniques were investigated for interpolating LM probabilities. Finally, English and Italian ASR systems were developed for the IWSLT 2014 evaluation.

Improvements were made to the Haystack multilingual multimedia information extraction and retrieval (MMIER) system that was initially developed under a prior work unit. Major additions to the user interface include the following: support for uploading multiple files, expansive changes to the media player, additional machine translation (MT) capabilities, and integration of geolocation information. Scripts were developed for translating web pages and displaying the results in the same format as the input. Research into HTML5 was initiated to improve functionality across different operating systems. The processing pipeline was updated to provide support for decoding hybrid DNN-HMM systems, support for N-gram and recurrent neural network (RNN) LM rescoring, and improved text extraction from portable document format (PDF) files. Japanese, Chinese, and Pashto speech recognition systems were developed and then incorporated into Haystack.

This report is organized as follows. Section 2.0 describes the experiments and accomplishments. Section 3.0 summarizes conclusions drawn from the experiments.

2.0 EXPERIMENTS AND ACCOMPLISHMENTS

This section discusses the experiments and accomplishments for the covered period. Section 2.1 discusses the ASR experiments that were performed, and Section 2.2 describes the improvements made to the Haystack MMIER system.

2.1 ASR Experiments

This section discusses the ASR experiments that were conducted. Section 2.1.1 describes how Korean ASR systems were designed to reduce the effects of OOV words. Section 2.1.2 presents the Levantine Arabic and Farsi ASR systems that were developed on conversational telephone speech. Section 2.1.3 describes three methods that were investigated for combining Pashto ASR systems. Section 2.1.4 discusses software that was developed for training and evaluating hybrid DNN-HMM speech recognition systems. Section 2.1.5 presents the English ASR system that was developed for the IWSLT 2013 evaluation campaign. Section 2.1.6 describes several methods that were investigated for performing LM interpolation. Finally, Section 2.1.7 describes the English and Italian ASR systems that were developed for IWSLT 2014.

2.1.1. Morfessor for Korean ASR

Korean ASR systems were designed to reduce the effects of OOV words encountered by the recognizer. OOV words are those words spoken by a person that are not in the pronunciation dictionary and LM for an ASR system; as a result, they will never appear in the output of the recognizer, thereby increasing the error rate. To reduce the number of OOV words, Korean LMs were estimated using both words and sub-word units that can be combined to form words.

Korean sub-word units were automatically derived using Morfessor [1] with the baseline algorithm and the categories-MAP algorithm with perplexity thresholds of 10, 50, 100, and 400. The following procedure was used to incorporate these sub-word units into the recognizer:

- Evaluate Morfessor on the text corpus
- Create a pronunciation dictionary by applying letter-to-sound rules
- Train an LM on the sub-word units, and attach a + sign to the start of every sub-word unit except for the first sub-word unit from a word
- Evaluate the recognizer using the pronunciation dictionary and sub-word LM
- Attach sub-word units that start with a + sign to the previous word or sub-word unit

This procedure was applied to text from GlobalPhone [2], the Korean Broadcast News corpus [3], the Korean Newswire corpus [4], and articles downloaded from Wikipedia.[1] Interpolated trigram LMs were estimated using the Stanford Research Institute LM (SRILM) toolkit [5]. Unless stated otherwise, all N-gram LMs discussed in this document were estimated using modified

[1] Available at: http://dumps.wikimedia.org/kowiki

Knesey-Ney smoothing. The vocabulary for each LM included 500000 tokens and was chosen using the select-vocab program from the SRILM toolkit.

Acoustic Models (AMs) were trained on GlobalPhone and the Korean Broadcast News corpus using HTK [6]. Pronunciations for all words were derived using letter-to-sound rules [7]. Phonemes were modeled using state-clustered across-word triphone HMMs, and the final HMM set included 3000 shared states with an average of 16 mixtures per state. The models were discriminatively trained using the Minimum Phone Error (MPE) criterion. The feature set consisted of 12 Perceptual Linear Prediction (PLP) coefficients, plus the zeroth coefficient, with mean normalization applied on a per utterance basis. Delta, acceleration, and third differential coefficients were appended to form a 52 dimensional feature vector, and Heteroscedastic Linear Discriminate Analysis (HLDA) was applied to reduce the feature dimension to 39. A second set of models was estimated that included Speaker Adaptive Training (SAT).

Each set of models was evaluated on the GlobalPhone and Korean Broadcast News development partition. Initial transcripts were produced using the HTK large vocabulary continuous speech recognizer HDecode. Constrained Maximum Likelihood Linear Regression (CMLLR) transforms were estimated for each speaker, and the final recognition hypotheses were generated using the SAT HMMs. Table 1 shows the Character Error Rate (CER) and Word Error Rate (WER) obtained with each system. The sub-word units yielded an improvement in CER and WER on both partitions.

Table 1: Korean CER and WER on the GlobalPhone and Korean Broadcast News Development Partitions

	GlobalPhone			**Broadcast News**	
Morfessor Algorithm	**CER**	**WER**		**CER**	**WER**
None	12.1	51.6		14.2	39.0
Baseline	10.9	42.9		13.4	38.5
Categories-MAP Perplexity 10	11.1	43.5		13.4	37.9
Categories-MAP Perplexity 50	11.3	45.6		13.5	38.5
Categories-MAP Perplexity 100	11.3	46.1		13.5	38.6
Categories-MAP Perplexity 400	11.3	47.2		13.5	38.7

2.1.2. Conversational Telephone ASR

This section describes the Levantine Arabic and Farsi ASR systems that were developed on conversational telephone speech. This is a particularly difficult task because conversational speech is highly coarticulated, less predictable than other types of speech (*e.g.*, read speech, lectures, or broadcast news), and typically includes more sentence restarts, word fragments, and filled pauses. In addition, ASR systems perform worse on telephone speech due to channel variability, reduced bandwidth, and transmission artifacts.

Levantine Arabic: An ASR system was developed on 31 hours of speech from the Levantine Arabic Conversational Telephone Speech Corpus (ARB-CTS) [8]. Prior to training the AMs, long periods of silence were removed from the audio files using an amplitude-based Speech Activity Detector (SAD). Gain normalization was applied to each utterance so that the maximum sample value was 32767 and 100 millisecond frames were extracted every 50 milliseconds.

Frames were classified as speech if the maximum sample value was greater than 2000, and silence otherwise. All speech end points were padded by 200 milliseconds and the silence regions were removed from each utterance. In a preliminary experiment, this process yielded a 7.5% absolute improvement in WER.

AMs were trained using the same procedure described in Section 2.1.1, except the feature mean and variance normalization were applied on a conversation side basis. The final HMM set included 3000 shared states with an average of 24 mixtures per state. The HMM system was evaluated using a trigram LM that was estimated on the training transcripts. This system yielded a 60.0% WER on the ARB-CTS test partition.

A second ASR system was developed on 138 hours of speech from the Levantine Arabic QT training data set 5 (ARB-QT) [9]. This system was trained using the same procedure described above, except that the amplitude SAD was not applied because the utterances did not include large regions of silence. The final HMM set included 5000 shared states with an average of 28 mixtures per state. Decoding was performed using a trigram LM that was estimated on the training transcripts. This system yielded a 51.0% WER on the ARB-QT test partition.

Farsi: Speech recognition systems were developed on eight hours of speech from the Appen mobile network mini database (ASR001) and 20 hours of speech from the the Appen conversational telephone speech corpus (ASR002).[2] Long periods of silence were removed from the training files using the amplitude based SAD described above. In a preliminary experiment, this yielded a 1.2% absolute improvement in WER.

An initial set of AMs was trained using the same procedure as the Levantine Arabic systems. The HMM system included 2000 shared states with an average of 20 mixtures per state. Phoneme alignments were generated for the entire training partition, and any utterance that included a phoneme duration greater than one second was sequestered from the training set. AMs were retrained on the modified training set using the same procedure described above. In a preliminary experiment, sequestering training utterances with long phoneme durations yielded a 0.7% absolute improvement in WER.

LMs were estimated on ASR002; the Translation System for Tactical Use (TRANSTAC) corpus; the Uppsala Persian corpus [10]; the Tehran English-Persian corpus [11]; translated text from Technology, Entertainment, And Design (TED) conferences;[3] and articles downloaded from Wikipedia.[4] Note that only the ASR002 text includes diacritics. One trigram LM was estimated on the ASR002 text that included diacritics, and a second LM was estimated on the same set of text with all diacritics removed. An interpolated trigram LM was trained on all of the text without diacritics.

Each system was evaluated on the ASR002 development partition, and all diacritics were removed prior to calculating the WER. The ASR002 LM with diacritics yielded a 60.6% WER, and the ASR002 LM without diacritics yielded a 62.6% WER. The interpolated trigram LM trained on all sources yielded a 62.3% WER.

[2] Appen corpora are available at: http://www.appen.com

[3] Available at: http://www.ted.com

[4] Available at: http://dumps.wikimedia.org/fawiki

2.1.3. Pashto System Combination

Three different methods were investigated for combining Pashto Speech Recognition Systems: Recognizer Output Voting Error Reduction (ROVER) [12], N-best ROVER, and word posterior decoding using matching scores from the Driven Decoding Algorithm (DDA) [13]. ROVER aligns the 1-best hypotheses from multiple ASR systems and applies a voting scheme to select the best transcript. The 1-best hypotheses were obtained using word posterior probability decoding [14], and ROVER was applied using the SRover program from the Brno toolkit.[5] N-best ROVER creates a confusion network using the N-best lists from multiple ASR systems and selects the word with the highest posterior probability from each correspondence set. This was accomplished using the nbest-rover program from the SRILM toolkit.

The third method computes matching scores by aligning N-best hypotheses from a primary ASR system to an auxiliary transcript produced by one or more secondary ASR systems. Each N-best hypothesis from the primary ASR system is aligned to the auxiliary transcript using a Dynamic Programming (DP) algorithm. The DP alignment was implemented using the same method as the sclite program from the National Institute of Standards and Technology (NIST) speech recognition scoring toolkit.[6] Consider a single hypothesis from the primary system $W = (w_1, w_2, \cdots, w_L)$ that is aligned to $W' = (w'_1, w'_2, \cdots, w'_L)$. the auxiliary transcript

A matching score $\theta(w_i)$ was assigned to each word based on the number of words in the history that match the auxiliary transcript

$$\theta(w_i) = \begin{cases} 0.99 & \text{if } \{w_j\} = \{w'_j\} \quad \text{for } i-3 \leq j \leq i \text{ and } j > 0 \\ 0.9 & \text{if } \{w_j\} = \{w'_j\} \quad \text{for } i-2 \leq j \leq i \text{ and } j > 0 \\ 0.4 & \text{if } \{w_j\} = \{w'_j\} \quad \text{for } i-1 \leq j \leq i \text{ and } j > 0 \\ 0.1 & w_i = w'_i \\ 0.01 & w_i \neq w'_i \end{cases} \tag{1}$$

The final score for each word was a weighted combination of the matching score, the AM score, the LM score, and the word insertion penalty. The 1-best transcript was selected from the N-best list using posterior probability decoding.

Each method was evaluated using three Pashto ASR systems: one hybrid DNN-HMM system and two HMM systems. The weights for each system were tuned on the TRANSTAC development partition using the nbest-optimize program from the SRILM toolkit. The hybrid DNN-HMM system was used as the primary system when calculating matching scores, and the auxiliary transcript was obtained by combining the two HMM system using N-best ROVER. Table 2 shows the WERs obtained on the TRANSTAC test partition. N-best ROVER yielded the best performance.

[5] Available at: http://speech.fit.vutbr.cz/software/hmm-toolkit-stk

[6] Available at: http://www.itl.nist.gov/iad/mid/tools

Table 2: Pashto WER on the TRANSTAC Test Partition

Three ASR systems were evaluated and system combination was performed using ROVER, N-best ROVER, and word posterior decoding with DDA matching scores.

System Combination	WER
None	34.4/33.4/32.9
ROVER	31.9
N-best ROVER	31.4
DDA matching scores	31.9

2.1.4. Hybrid DNN-HMM Systems

This section describes the software that was developed for training and evaluating hybrid DNN-HMM speech recognition systems. Whereas standard HMM systems model observation probabilities using Gaussian Mixture Models (GMMs), hybrid DNN-HMM systems replace the GMMs in a well-trained HMM system with a DNN. In the context of this paper, DNNs are feed forward neural networks with more than one hidden layer. The procedure for developing a hybrid DNN-HMM system can be summarized as follows:

- Train a state-clustered GMM-HMM system
- Generate HMM state-level time alignments of the training data using forced alignment
- Train a DNN to model the shared states of the GMM-HMM system
- Use the DNN instead of the GMMs when evaluating the recognizer

The GMM-HMM system and state-level time alignments can be generated using HTK. DNNs were trained using layer growing back propagation [15]. This method estimates the parameters for a DNN by first initializing a one hidden layer network with random weights and training the network to convergence using error back propagation. Next, the output layer is replaced with a second randomly initialized hidden layer, followed by a randomly initialized output layer. This network is then trained to convergence, and the process of replacing the output layer and retraining the network is repeated until the DNN includes the desired number of hidden layers.

Two different programs were investigated for training DNNs: the International Computer Science Institute (ICSI) QuickNet software package[7] and Theano [16]. To train DNNs with QuickNet, software was developed to convert HTK state-level time alignments to QuickNet pfile format and to replace the output layer in QuickNet Matlab Level-4 network files with a randomly initialized hidden layer and output layer. One limitation of QuickNet is that it only supports a maximum of three hidden layers; software was developed using Theano to train deeper networks. Python code was written to read input vectors into a cache, apply a context window, remove unwanted samples, randomize the data, and copy the data to the Graphical Processing Unit (GPU). The DNN training algorithm and evaluation routines were implemented in Theano by modifying the multilayer perceptron code from [17].

[7] Available at: http://www1.icsi.berkeley.edu/Speech/icsi-speech-tools.html

Table 3: English WER on the IWSLT dev2010 Partition using Hybrid DNN-HMM Systems

QuickNet was used to train DNNs with 1–3 hidden layers, and Theano was used to train DNNs with 1–5 hidden layers.

	Hidden Layers				
DNN software	1	2	3	4	5
QuickNet	24.	21.	20.	–	–
Theano	24.5	21.7	20.7	20.1	19.8

Lastly, HDecode and the Sphinx-4 speech recognizer[8] were modified to read HMM state likelihoods from HTK feature files [18]. For a given state s and observation vector o, the posteriors from the DNN were converted to likelihoods by dividing by the prior probability of each state, *i.e.*,

$$P(o|s) = \frac{P(s|o)P(o)}{P(s)}, \tag{2}$$

where $P(s|o)$ is the posterior probability estimated by the DNN, $P(s)$ is the prior probability of s estimated from the training data, and $P(o)$ is a constant that can be ignored.

To compare QuickNet and Theano, hybrid DNN-HMM systems were developed on 58 hours of TED talks. The GMM-HMM models were trained using the same procedure described in Section 2.1.1, and the final HMM set included 3000 shared states with an average of 24 mixtures per state. DNNs were trained using a maximum of 5 hidden layers, each of which had 1000 neurons with logistic activation functions. A context window of 9 frames was used at the input, and the output included 3000 units corresponding to the shared states of the GMM-HMM system. The feature set consisted of 13 PLPs with delta and acceleration coefficients, and all features were normalized to zero mean and unit variance on a per speaker basis. Training was performed with a minibatch size of 512, and an initial learning rate of 0.008 that was halved after each epoch once the improvement in accuracy on the cross validation partition fell below 0.5%. Training was completed once the improvement in accuracy fell below 0.5% a second time.[9]

Each system was evaluated on the dev2010 partition from the IWSLT evaluation campaign [19]. Decoding was performed using a single pass of HDecode with a trigram LM that was developed for IWSLT 2012 [20]. Table 3 shows the WERs obtained with each DNN. For comparison purposes, the GMM-HMM system was evaluated using the same procedure described in Section 2.1.1; the first pass yielded a 22.0% WER, and the second pass yielded a 19.9% WER.

2.1.5. IWSLT 2013

This section describes the English ASR system that was developed for the IWSLT 2013 evaluation campaign. This task focuses on the automatic transcription of TED talks, which are

[8] Available at: http://cmusphinx.sourceforge.net

[9] This is the QuickNet newbob training strategy

professionally recorded presentations given on a variety of topics related to technology, entertainment, and design.

Each talk is a maximum of 18 minutes in length. The TED website[10] makes the video recordings and closed captions from over 1900 talks available for download.

AMs were trained on 807 TED talks that were recorded prior to 2011. The audio was extracted from each video file using FFmpeg,[11] and then downsampled to 16 kHz using SoX.[12] Long periods of untranscribed audio were removed from each talk using the time marks from the closed captions, and word alignments were automatically generated using an HTK HMM system developed on HUB4 [21, 22]. These alignments were used to split each talk into utterances that were shorter than 20 seconds and included 0.1– 0.25 seconds of non-speech at the end points. Next, closed caption filtering [23] was applied to the TED data to sequester utterances that may include transcription errors. Each talk was decoded using the HUB4 HMMs and a trigram LM that was estimated on the transcripts for the talk. The recognizer outputs were compared to the transcripts, and a data partition was created using all utterances with a WER less than 30%. This process yielded 166 hours of audio.

A speaker independent hybrid DNN-HMM speech recognition system was developed using the Theano software described in Section 2.1.4. The GMM-HMM system included 6000 shared states with an average of 28 mixtures per state; the DNN included a context window of 9 frames on the input, 5 hidden layers with 1000 units each, and 6000 output units. A speaker adaptive DNN was trained on PLP features that were transformed using CMLLR. This system applied a single transform per speaker.

LMs were developed on the TED data provided by IWSLT,[13] the English Gigaword corpus [24], and the News 2007–2012 texts from the Association for Computational Linguistics Workshop on Machine Translation (WMT).[14] Cross entropy difference scoring [25] was used to select subsets of Gigaword and News 2007–2012 that matched the TED domain. Interpolated trigram and 4-gram LMs were estimated on TED, 1/8 of Gigaword, and 1/4 of News 2007–2012. RNN maximum entropy LMs were developed using the RNNLM toolkit [26]. One RNN was trained on 1/16 of Gigaword, and a second RNN was trained on 1/8 of News 2007–2012. Each network included 160 hidden units, 300 classes in the output layer, 4-gram features for the direct connections, and a hash size of 10.[9] The LM vocabulary included 95000 words and was chosen using the select-vocab program from the SRILM toolkit.

Whereas in previous IWSLT evaluations [19, 27] the test data was manually segmented into spoken utterances, this year each talk was provided without timing information. A neural network-based SAD was developed using Theano to segment each talk into utterances and remove long periods of non-speech. The SAD was trained on 22 hours of TED data and 5 hours of public domain music downloaded from Wikimedia Commons,[15] the United States Air Force band,[16] and the Open Goldberg Variations project.[17] The network included a context window of 21 frames on the input, 1 hidden layer of 500 neurons with logistic activation functions, and 3

[10] http://www.ted.com

[11] Available at: http://www.ffmpeg.org

[12] Available at: http://sox.sourceforge.net

[13] Available at: http://workshop2013.iwslt.org

[14] Available at: http://www.statmt.org/wmt13/translation-task.html

[15] Available at: http://commons.wikimedia.org

[16] Available at: http://www.usafband.af.mil

[17] Available at: http://www.opengoldbergvariations.org

Table 4: English WER on the IWSLT 2012 Development Partitions using Manual and Automatic Segmentations of the Data

	Manual				Automatic		
System	**dev2010**	**tst2010**	**dev2012**		**dev2010**	**tst2010**	**dev2012**
Decode-1	14.3	13.0	15.3		15.6	14.3	16.9
Decode-2	13.7	12.3	14.0		14.8	13.5	15.8
4-gram	13.1	11.6	13.2		13.9	12.7	14.9
4-gram + RNN	12.1	10.3	11.6		12.8	11.8	13.8

output units corresponding to speech, silence/noise, and music. The feature set consisted of 12 PLP coefficients, plus the zeroth coefficient, with delta and acceleration coefficients. All features were globally normalized to zero mean and unit variance. Six epochs of training were performed with a minibatch size of 512, and an initial learning of 0.008 that was halved after the second epoch.

Automatic segmentation of the test data was performed by evaluating the SAD, applying a DP algorithm to choose the best sequence of states, and padding the speech end points by 0.15 seconds. The speech segments from each talk were clustered using the Massachusetts Institute of Technology Lincoln Laboratory (MIT-LL) GMM software package [28]. Initial transcripts of the test data were produced using HDecode with the interpolated trigram LM. These transcripts were used to estimate CMLLR transforms for the speaker adaptive hybrid DNN-HMM system. A second pass of HDecode was evaluated to generate recognition lattices, which were then rescored with the interpolated 4-gram LM. Next, 1000-best lists were extracted from each lattice and rescored with the RNN LMs. The final LM scores were obtained by linearly interpolating the probabilities from the 4-gram and RNN LMs. Lastly, the maximum scoring utterance was selected for each utterance.

Table 4 shows the WERs obtained on the IWSLT development partitions at each decoding stage. For comparison purposes, results are shown on both the manually produced and automatically derived segmentations of the data. This system yielded a 15.9% WER on the tst2013 partition and placed third out of the eight ASR systems that were submitted for the evaluation.

2.1.5 LM Interpolation

Six different methods were investigated for interpolating probabilities from 4-gram and RNN LMs. Note that the LMs described in this paper estimate the probability $P\ (w|h)$ for a word w with history h, where $0 \leq P\ (w|h) \leq 1$. One of the most popular methods for combining probabilites from multiple models is linear interpolation. Given the probabilities $Pk\ (w|h)$ from N models, the interpolated probability can be calculated as

$$P(w|h) = \sum_{k=1}^{N} \lambda_k P_k(w|h), \tag{3}$$

where λ_k is the interpolation weight for the k_{th} model. The interpolation weights are typically subject to the constraints

$$0 \leq \lambda_k \leq 1 \qquad \text{and} \qquad \sum_{k=1}^{N} \lambda_k = 1.$$

A modified version of linear interpolation was implemented where the range of probabilities $P_k(w|h)$ from the individual models was restricted. For a given word w and history h, the minimum probability from any model was set to the maximum probability divided by an empirically chosen integer L, that is

$$\begin{aligned} P_{min} &= \left(\max_{1 \leq k \leq N} P_k(w|h) \right) / L, \\ P_k(w|h) &= P_{min} \qquad \text{if } P_k(w|h) < P_{min}. \end{aligned} \tag{4}$$

Recall that Equation 3 computes a weighted sum. Alternatively, $P\ (w|h)$ was calculated by selecting the minimum, maximum, or median of $P_k\ (w|h)$ for $1 \leq k \leq N$. Lastly, LM interpolation was performed by linearly interpolating the log probabilities from multiple models

$$P(w|h) = \exp \sum_{k=1}^{N} \lambda_k \log P_k(w|h). \tag{5}$$

Each method of LM interpolation was evaluated on the IWSLT 2013 development partitions. The probabilites $P_k\ (w|h)$ were obtained from the ASR system described in Section 2.1.5. This system provided 1000-best lists that were scored with three different models: one 4-gram LM and two RNN LMs. These three LMs are referred to as *forward* models in the remainder of this section. A second set of *backward* RNN LMs were developed on Gigaword and News 2007–2012. These models were trained using the same procedure described in Section 2.1.5, except that the word order of the input text was reversed during training and evaluation. The backward RNN LMs were used to rescore the same set of 1000-best lists.

The linear interpolation method described by Equation 4 was evaluated using L = 5, 10, 20, 100 and the interpolations weights λ_k were optimized using the compute-best-mix program from the SRILM toolkit. Each interpolation method was evaluated using two different sets of LMs: the first set included the three forward LMs, and the second set included the three forward LMs and two backward LMs. Table 5 shows the WERs obtained. Linearly interpolating the log probabilities from each model yielded the best results, especially when including the backward LMs.

2.1.6. IWSLT 2014

English and Italian ASR systems were developed for the IWSLT 2014 evaluation campaign. This task focuses on the automatic transcription of English TED talks and Italian TEDx talks. TEDx talks are similar to TED, but given on a wider array of topics at independently organized events across the world. Whereas TED talks typcially include high quality speech, TEDx talks are recorded with varying degrees of quality and may include reverberated speech, background noise, or audio compression artifacts.

Table 5: English WER on the IWSLT Development Partitions using Six Different Methods for Interpolating Probabilities from 4-gram and RNN LMs

Each method was evaluated using (1) the three forward LMs and (2) the three forward LMs and two backward LMs.

	Forward LMs				Forward and Backward LMs		
Interpolation	**dev2010**	**tst2010**	**dev2012**		**dev2010**	**tst2010**	**dev2012**
Linear	12.1	10.3	11.6		13.6	12.4	14.0
Linear L=5	12.2	10.4	11.7		13.9	12.8	14.5
Linear L=10	12.1	10.3	11.6		13.8	12.5	14.2
Linear L=20	12.1	10.3	11.6		13.6	12.4	14.1
Linear L=100	12.1	10.3	11.6		13.6	12.4	14.0
Linear maximum	12.4	10.9	12.2		14.7	13.9	15.5
Linear minimum	12.4	10.7	12.2		13.9	12.4	13.9
Linear median	12.3	10.4	11.9		12.2	11.1	12.0
Log linear	11.8	10.0	11.6		11.7	9.9	11.4

English: In addition to the TED acoustic data described in Section 2.1.5, AMs were trained on broadcast news speech from the HUB4 and Euronews [29] corpora. The audio from each corpus was segmented into utterances using the manually produced transcripts for HUB4 and the provided ASR transcripts for Euronews. All utterances were processed with a GMM-based bandwidth detector to identify and remove telephone bandwidth speech. The MIT-LL GMM software package was used to automatically cluster utterances from the Euronews corpus. This process yielded 128 hours of audio from HUB4 and 96 hours from Euronews.

An HMM system was trained on TED using the same procedure described in Section 2.1.1, except that feature mean and variance normalization was applied on per speaker basis. The final HMM set included 6000 shared states with an average of 28 mixtures per state. Hybrid DNN-HMM speech recognition systems were developed on TED, HUB4, and Euronews using the same procedure described in Section 2.1.5. The GMM-HMM set included 8000 shared states with an average of 28 mixtures per state; each DNN included a context window of 9 frames in the input, 7 hidden layers with 1000 units each, and 8000 output units.

LMs were developed on the TED data provided by IWSLT,[18] the English Gigaword corpus, and the News 2007–2013 texts from WMT.[19] Data selection was implemented using the same procedure described in Section 2.1.5. Interpolated trigram and 4-gram LMs were estimated on TED, 1/8 of Gigaword, and 1/8 of News 2007–2013. An RNN maximum entropy LM was trained on the same set of training texts using the RNNLM toolkit. The network included 160 hidden units, 300 classes in the output layer, 4-gram features for the direct connections, and a hash size of 10_9. The LM vocabulary included 100000 words and was chosen using the select-vocab program from the SRILM toolkit.

[18] Available at: http://workshop2014.iwslt.org

[19] Available at: http://www.statmt.org/wmt14/translation-task.html

Table 6: English WER on the IWSLT 2014 Development Partitions

Results are given for the HMM/DNN-HMM systems when both systems were evaluated in parallel.

System	dev2010	tst2010	dev2012
Decode-1	14.8	13.4	16.2
Decode-2	14.6/14.3	12.7/12.8	15.3/14.8
4-gram	14.0/13.7	12.3/12.1	14.6/14.2
4-gram + RNN	13.0/12.6	11.5/11.6	13.7/13.3
N-best ROVER	11.6	10.4	12.4

The decoding procedure is shown in Figure 1. Automatic segmentation of the test data was performed using the same procedure described in Section 2.1.5, and initial transcripts were produced using HDecode with the speaker independent hybrid DNN-HMM system and the trigram LM. The HMM system and the speaker adaptive hybrid DNN-HMM system were then evaluated in parallel using the following decoding strategy. First, the initial transcripts were used to estimate CMLLR feature transforms for each speaker. Next, recognition lattices were generated using Sphinx-4 with the HMM system and HDecode with the speaker adaptive hybrid DNN-HMM system. The lattices were rescored with the interpolated 4-gram LM, and 1000-best lists were extracted from each lattice for rescoring with the RNN LM. The final LM scores were obtained by linearly interpolating the log probabilities from the 4-gram and RNN LM. Lastly, system combination was performed using the N-best ROVER program from the SRILM toolkit.

Table 6 shows the WERs obtained on the IWSLT development partitions at each decoding stage. The final submission to the IWSLT evaluation included an additional Tandem ASR system that was developed by MIT-LL [30]. The recognition lattices from this system were rescored using the same procedure described above, and the outputs from all three systems were combined using N-best ROVER. The final system yielded a 9.9% WER on the tst2014 partition and placed third out of the eight ASR systems that were submitted for the evaluation.

Italian: An Italian pronunciation dictionary was manually created for the most frequent 28000 words from the Euronews corpus. This was done by a member of the Speech and Communication Research, Engineering, Analysis, and Modeling (SCREAM) laboratory who speaks Italian as a second language.[20] The 51 phone set included 24 non-geminated consonants, 20 geminated consonants, and 7 vowels. The consonants M, N, j, w, z were never geminated and the consonant ñ was always geminated. A second pronunciation dictionary with 32 phones was created by ignoring gemination. Lastly, a multilingual pronunciation dictionary was created from the Italian dictionary that ignored gemination and version 0.7a of the English Carnegie Mellon University (CMU) pronunciation dictionary.[21] Italian and English phones were merged when they shared the same International Phonetic Alphabet (IPA) symbol. Table 7 shows the phone set for each language; the English phones are in ARPAbet format.[22] The multilingual dictionary included 48 phones.

[20] Thanks to Kyle Wilkinson for creating the Italian pronunciation dictionary

[21] Available at: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[22] The ARPAbet to IPA mappings used in this work are available at: http://en.wikipedia.org/wiki/Arpabet

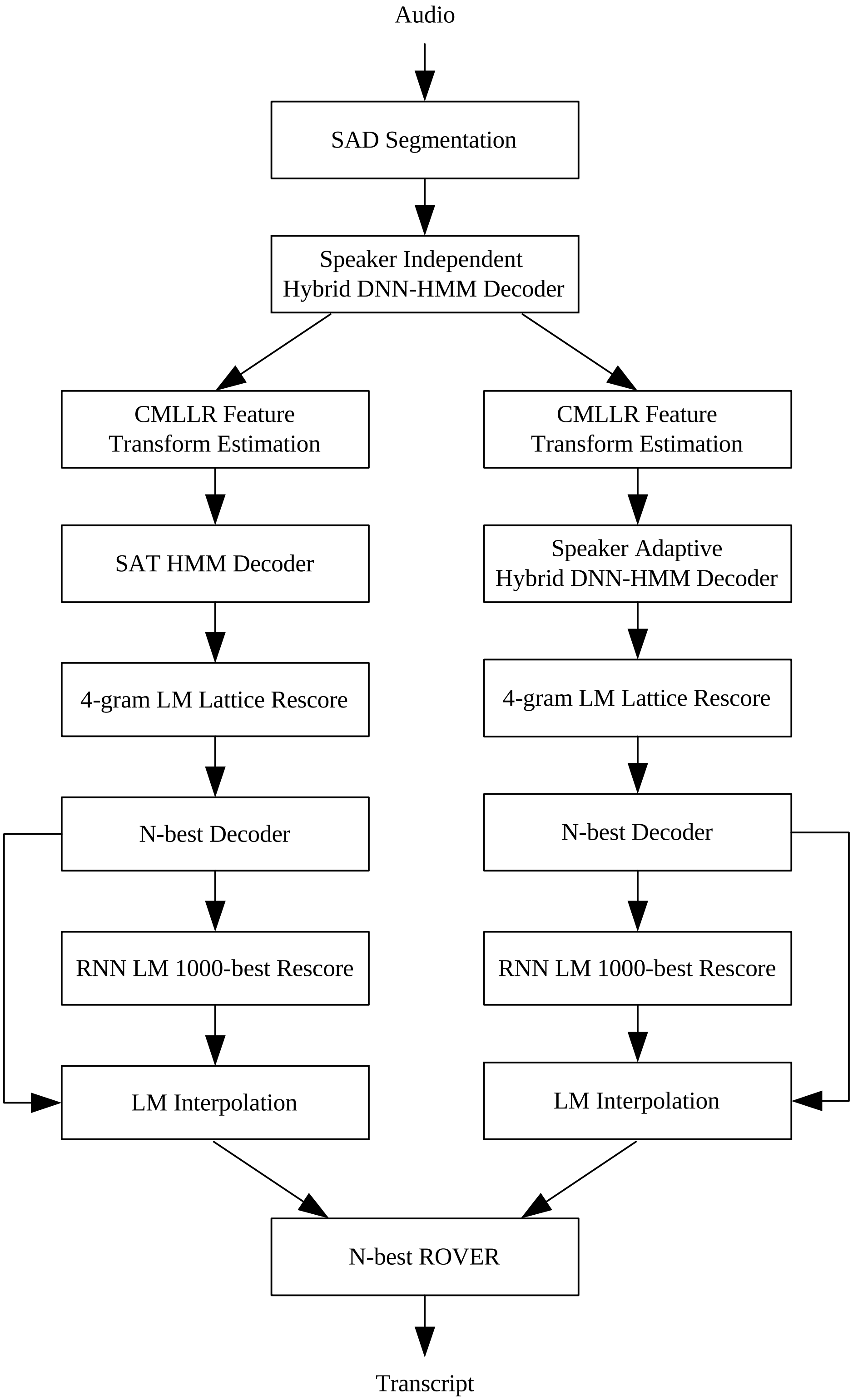


Figure 1: IWSLT 2014 English Decoding Procedure

Table 7: Italian and English Phone Sets

Dashes indicate that a phone does not exist in the corresponding language.

IPA	Italian	English	IPA	Italian	English	IPA	Italian	English
p	P	P	s	S	S	a	A	–
b	B	B	z	Z	Z	u	UW	UW
t	T	T	S	SH	SH	o	O	–
d	D	D	Z	–	ZH	0	AO	AO
k	K	K	h	–	HH	A	–	AA
g	G	G	t͡s	TS	–	æ	–	AE
m	M	M	d͡z	DZ	–	2	–	AH
M	EM	–	t͡S	CH	CH	@	–	AX
n	N	N	d͡Z	JH	JH	Ç	–	ER
ñ	NY	–	j	Y	Y	I	–	IH
N	NG	NG	l	L	L	U	–	UH
r	R	R	L	GL	–	aU	–	AW
f	F	F	w	W	W	aI	–	AY
v	V	V	i	IY	IY	eI	–	EY
T	–	TH	E	EH	EH	oU	–	OW
D	–	DH	e	E	–	0I	–	OY

HMM and hybrid DNN-HMM systems were trained on the Euronews Italian data set using the same procedure as the English systems. One HMM system was trained using the 51 phone set (denoted as HMM-51), and a second HMM system was trained using the 32 phone set (denoted as HMM-32). HMM-51 included 6000 shared states with an average of 28 mixtures per state, and HMM-32 included 4000 shared states with an average of 24 mixtures per state. The hybrid DNN-HMM system was developed using HMM-51; the DNNs included 3 hidden layers with 1000 units each and 6000 output units. A final HMM system (denoted as HMM-ML) was developed on Euronews Italian and English TED using the multilingual pronunciation dictionary; HMM-ML included 6000 shared states with an average of 28 mixtures per state.

Interpolated trigram and 4-gram LMs were developed on the TED data provided by IWSLT,[23] the Google Books Ngram corpus [31], and the Web 1T 5-gram corpus [32]. Words from the TED data set were split on apostrophes, and N-grams from Google Books were ignored if the source was published prior to the year 2000. The TED LMs were estimated using modified Kneser-Ney smoothing; the Google Books and Web 1T LMs were estimated using Witten-Bell smoothing. An RNN maximum entropy LM was trained on TED; the network included 320 hidden units, 200 classes in the output layer, 4-gram features for the direct connections, and a hash size of 10^9. The LM vocabulary included 100000 words and was chosen using the select-vocab tool from the SRILM toolkit.

[23] Available at: http://workshop2014.iwslt.org

Automatic segmentation of the test data was initially performed using the same procedure described in Section 2.1.5. On the dev2014 partition, it was discovered that the SAD was misclassifying non-speech segments as speech on several TEDx talks. To alleviate this problem, any speech segment longer than 20 seconds was reprocessed with a previously developed neural network based SAD. This SAD was created using QuickNet and trained on English telephone speech from the Fisher corpus [33]. The network included a context window of 9 frames on the input, 1 hidden layer of 1400 units with logistic activation functions, and 4 output units corresponding to voiced speech, unvoiced speech, aspirated speech, and non-speech. The feature set consisted of 12 PLP coefficients, plus energy, with delta and acceleration coefficients. All features were globally normalized to zero mean and unit variance. As with the English system, speech segments from each talk were clustered using the MIT-LL GMM software package.

Decoding was performed as follows. Initial transcripts of the test data were produced using HDecode with the speaker independent hybrid DNN-HMM system and the trigram LM. The HMM-32, HMM-ML, and speaker adaptive hybrid DNN-HMM systems were then evaluated in parallel using HDecode with the same decoding strategy as the English system. Finally, system combination was performed using N-best ROVER.

It was discovered that there were a number of errors in the reference transcripts for the IWSLT dev2014 partition. Therefore, a member of the SCREAM laboratory[24] manually corrected the reference transcripts for all 13 TEDx talks. Table 8 shows the WERs on the dev2014 partition at each decoding stage. For comparison purposes, results are included without cross adaptation of the HMM-32 and HMM-ML systems; that is, each system was evaluated independently instead of using the initial transcripts from the speaker independent hybrid DNN-HMM system. Results are also included when N-best ROVER was also applied at each decoding stage. From Table 8 we can see that cross adaptation of the HMM-32 and HMM-ML systems improved the WER. The final system yielded a 23.0% WER on the tst2014 partition and placed second out of the four ASR systems that were submitted for the evaluation.

2.2 Haystack MMIER System

This section describes improvements made to the Haystack MMIER system. Section 2.2.1 discusses improvements made to the user interface. Section 2.2.2 discusses several improvements that were made to the processing pipeline. Section 2.2.3 describes the Japanese, Chinese, and Pashto ASR systems that were developed for Haystack.

2.2.1. User Interface Improvements

Recent work in Haystack has seen a growth from version 0.6 to 0.8. There have been many additions to the user interface, including multiple file upload abilities, expansive changes to the Haystack Media Player, additional MT capabilities, and research into geolocation. To expand the toolset of Haystack there has been the testing of Optical Character Recognition (OCR) as a new avenue for media translation and the development of scripts to allow for complete webpages to be uploaded for translation but keeping the format intact. Future growth of Haystack includes cutting ties to tools that limit its functionality across the spectrum of operating systems, such

[24] Thanks to Kyle Wilkinson for correcting the transcripts

Table 8: Italian WER on the IWSLT dev2014 Partition

The HMM-32 and HMM-ML systems were evaluated both with and without cross adaptation. N-best ROVER was applied at each decoding stage. WER was calculated using (a) the provided reference transcripts and (b) the corrected reference transcripts.

(a) Provided Reference Transcripts

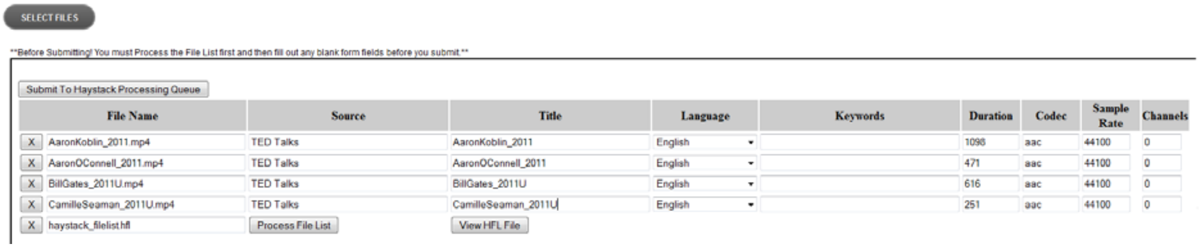
System	Decode-1	Decode-2	4-gram	4-gram + RNN
No Cross Adaptation				
DNN-HMM	35.0	32.9	32.5	32.5
HMM-32	41.2	34.4	34.1	33.9
HMM-ML	42.7	35.9	35.7	35.4
N-best ROVER	35.2	31.3	30.8	30.8
With Cross Adaptation				
DNN-HMM	35.0	32.9	32.5	32.5
HMM-32	–	32.2	31.8	31.4
HMM-ML	–	32.4	32.3	32.3
N-best ROVER	–	30.1	29.7	29.5

(b) Corrected Reference Transcripts

System	Decode-1	Decode-2	4-gram	4-gram + RNN
No Cross Adaptation	30.7	27.9	27.6	27.8
HMM-32	37.3	29.8	29.4	29.4
HMM-ML	39.1	31.3	31.0	30.9
N-best ROVER	31.4	26.7	26.3	26.4
With Cross Adaptation				
DNN-HMM	30.7	27.9	27.6	27.8
HMM-32	–	27.3	27.0	26.6
HMM-ML	–	27.5	27.5	27.5
N-best ROVER	–	25.3	25.0	25.0

as Adobe Flash, so research into HTML5 was initiated.

Multiple File Upload: There was a need for uploading multiple files into the Haystack service. A Flash application was developed for opening a file directory window and allowing for multiple file selection for upload. Once the files are uploaded, an interface is created to display the uploaded files in the queue, and each file is automatically processed with a metadata scan by FFmpeg for duration, codec, sample rate, etc. Each file also has form fields that can be populated with file information, such as the source, title, and source language.



Before Submitting! You must Process the File List first and then fill out any blank form fields before you submit.

	File Name	Source	Title	Language	Keywords	Duration	Codec	Sample Rate	Channels
X	AaronKoblin_2011.mp4	TED Talks	AaronKoblin_2011	English		1098	aac	44100	0
X	AaronOConnell_2011.mp4	TED Talks	AaronOConnell_2011	English		471	aac	44100	0
X	BillGates_2011U.mp4	TED Talks	BillGates_2011U	English		616	aac	44100	0
X	CamilleSeaman_2011U.mp4	TED Talks	CamilleSeaman_2011U	English		251	aac	44100	0
X	haystack_filelist.hfl	Process File List	View HFL File						

Figure 2: The Multiple File Upload Application after Processing an HFL File

For ease in uploading files across servers and from various directories, a Haystack File List (HFL) format was developed. This tab-separated text file can be created that includes file location, source, title, language, and keywords for multiple file data; it can be uploaded to the system and processed all at once. Figure 2 shows a screenshot of the multiple file upload application after processing an HFL file.

Geolocation: GeoNames[25] is a creative commons geographical database with over 10 million geographical names integrated with geographical data such as population, elevations, and latitude/longitude coordinates. After reconfiguring the Solr schema, the GeoNames database was indexed and experiments begun on linking the geographical coordinates to named entities identified by Janya in English and Chinese. A new search interface was also created to allow for searching for Haystack entries containing geographical locations or anything within a specific latitude/longitude distance.

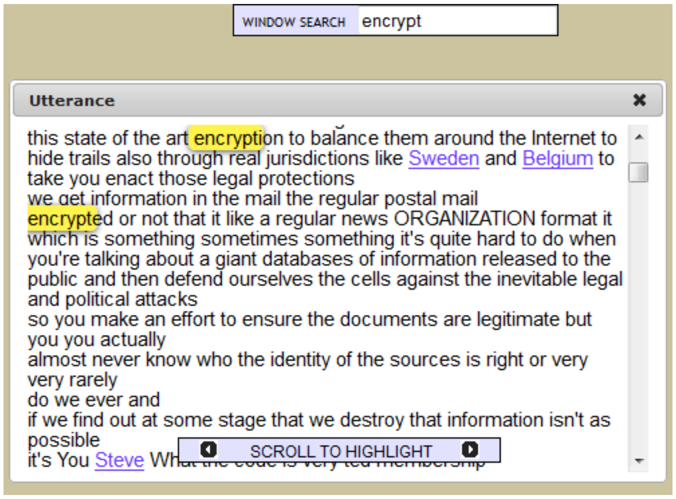
Relevancy was an issue because of the nature of redundant names in locations throughout the world so some methods were implemented to improve the reliability. Lists of population and popularity were created to check against results from geographical queries, and scripts were developed to add weight to the results in favor of those lists.

The next step involved integrating the OpenLayers[26] library for displaying map data into Haystack. Scripts were developed for integrating the list of locations into a tab-based interface and displaying the location markers on the map with links to Wikipedia entries.

Media Player: The Media Player section of Haystack has been subjected to constant updates. JQuery[27] opened up many new options in easing operation and customization for the user. A more graceful tab-based system was created across the top of the page to allow for easy access to utterances, translations by MT engine, speaker, topic, file metadata, and geographical functions. Also accessible through the tabs are a link to the auxiliary file data used in the pipeline process for Haystack and a link to a viewer for the log file.

The code was rewritten to update the dynamic modal windows for captioning. This update allows for smoother opening and closing of windows and better control of dragging the windows for placement or resizing with corner click-dragging. By fine tuning the window controls, it is now

[25] Available at: http://www.geonames.org
[26] Available at: http://openlayers.org
[27] Available at: http://jquery.com



WINDOW SEARCH encrypt

this state of the art encryption to balance them around the Internet to hide trails also through real jurisdictions like Sweden and Belgium to take you enact those legal protections
we get information in the mail the regular postal mail
encrypted or not that it like a regular news ORGANIZATION format it which is something sometimes something it's quite hard to do when you're talking about a giant databases of information released to the public and then defend ourselves the cells against the inevitable legal and political attacks
so you make an effort to ensure the documents are legitimate but you you actually
almost never know who the identity of the sources is right or very very rarely
do we ever and
if we find out at some stage that we destroy that information isn't as possible
it's You Steve What the code is very ted membership

Figure 3: Highlight and Scrolling Functionality Active in an Utterance Window

standard that when a play command is given from the search results whichever MT engine chosen is the default window open in the Player.

The viewer for of processed text files has been updated so that all of the translated text is parallel with the source material and each MT engine output is tab-based for viewing one at a time or all at once.

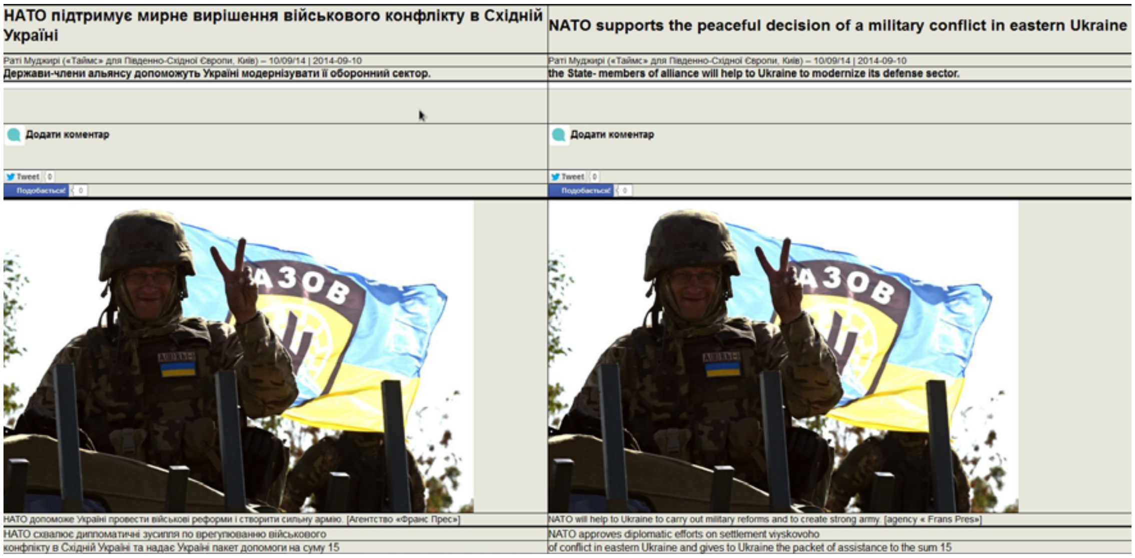
Highlighting: In Haystack, Solr is used for querying the vast index of processed media files, but once within the Media Player section of a specific file, a page-centric search functionality was created that would highlight every occurrence of those results.

This within page search allows a user to input a search term and see that term highlighted in each available window. Each window, in turn, has its own controller for scrolling back and forth between each highlighted term and can begin playing the file from that point. Figure 3 shows a screenshot of the hightlight and scrolling functionality active in an Utterance window.

HTML5: The initial file upload and media viewing capabilities of Haystack were Flash-based applications. As HTML5 evolved and browsers began to adopt its functionality, research began on how it might be leveraged to replace the Flash-based tools within the Haystack system. A rudimentary player was developed using HTML5, with simple controls and limited captioning options. To allow for cross-browser compatibility of the player, functionality was added to the pipeline to covert audio and video files into OGG and MP4 formats.

A new File Upload system was developed in HTML5 that allows for thumbnail viewing and playing and shows upload progress. The system allows for multiple file select but from only a standard file browser window, limiting the capabilities available through the HFL file option in the original Multiple File Upload system. Research was conducted on making the File Upload system more robust.

HTML Conversion: One missing factor in Haystack was the ability to upload or point to a webpage address and upload it for translation. A technique was developed using JavaScript and the Document Object Module (DOM) for place keeping of the text and images so that post-translation they could be placed back in the correct format.



НАТО підтримує мирне вирішення військового конфлікту в Східній Україні	NATO supports the peaceful decision of a military conflict in eastern Ukraine
Раті Муджирі («Таймс» для Південно-Східної Європи, Київ) – 10/09/14 | 2014-09-10	Раті Муджирі («Таймс» для Південно-Східної Європи, Київ) – 10/09/14 | 2014-09-10
Держави-члени альянсу допоможуть Україні модернізувати її оборонний сектор.	the State- members of alliance will help to Ukraine to modernize its defense sector.
НАТО допоможе Україні провести військові реформи і створити сильну армію. [Агентство «Франс Прес»]	NATO will help to Ukraine to carry out military reforms and to create strong army. [agency « Frans Pres»]
НАТО схвалює дипломатичні зусилля по врегулюванню військового	NATO approves diplomatic efforts on settlement viyskovoho
конфлікту в Східній Україні та надає Україні пакет допомоги на суму 15	of conflict in eastern Ukraine and gives to Ukraine the packet of assistance to the sum 15

Figure 4: The Output from the HTML Conversion and Translation Tools

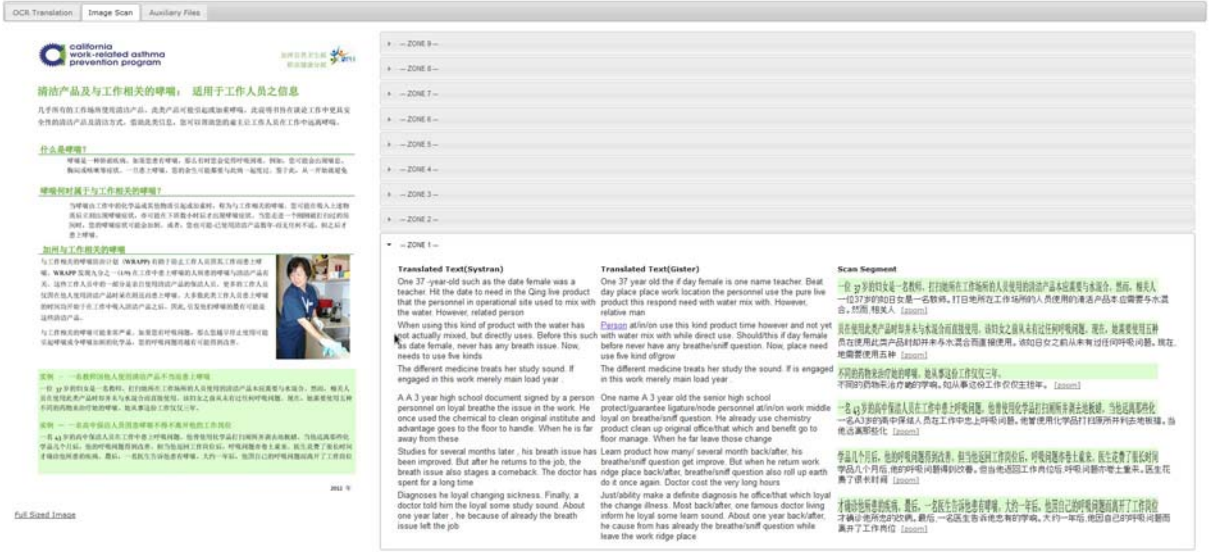
A prototype was created that parsed the HTML and sent only the text on for translation. The returned text was integrated back into the HTML framework and the results can be viewed side-by-side. For the prototype, only the Systran7 MT engine was used. Figure 4 shows a screenshot of the HTML conversion application. Development is continuing to make the HTML uploader a fully functioning part of the Haystack system.

OCR: The SCREAM Lab received a copy of the Raytheon/BBN Document Analysis Service (DAS) to test out as an OCR option to use in Haystack. The default language system packaged with it was Chinese.

Research began on creating a pipeline to tie into the DAS system and to optimize the input and output for best translation. After considerable testing, an image resolution of 400 dots per inch (DPI) was considered optimal. The initial phase allows for an image to be uploaded into the Haystack system, but the second stage is prompted by a command line instruction to the DAS system itself to commence the OCR operation and output the Extensible Markup Language (XML-based files back into the Haystack directory structure. A third phase is then initiated to translate the extracted text and place it back into the framework of a newly created viewer in the Media Player section.

Continued development of the system for the Chinese DAS OCR image viewer allows for tabs to see extracted text, translated text, and scanned zones of the image—also allowing for clicking on *zones* within the main image to scroll to the translation/OCR segment. A *zoom* function was developed so that scanned segments could be viewed at a greater magnification to check against the OCR output. Figure 5 shows the prototype output of intergrating OCR into Haystack.

Machine Translation: A pipeline was developed to integrate the Moses machine translation server into Haystack. Systems were integrated for French, Spanish, German, Farsi, Pashto, Arabic and to normalize, tokenize, and recase the text



Translated Text(Systran)	Translated Text(Gister)	Scan Segment
One 37 -year-old such as the date female was a teacher. Hit the date to need in the Qing live product that the personnel in operational site used to mix with the water. However, related person	One 37 year old the if day female is one name teacher. Beat day place place work location the personnel use the pure live product this respond need with water mix with. However, relative man	一位37岁的如日女是一名教师。打日地所在工作场所的人员使用的清洁产品本应需要与水混合。然而，相关人 [zoom]
When using this kind of product with the water has not actually mixed, but directly uses. Before this such as date female, never has any breath issue. Now, needs to use five kinds	Person at/in/on use this kind product time however and not yet with water mix with while direct use. Should/this if day female before never have any breathe/sniff question. Now, place need use five kind of/grow	员在使用此类产品时却并未与水混合而直接使用。该如日女之前从未有过任何呼吸问题。现在，她需要使用五种 [zoom]
The different medicine treats her study sound. If engaged in this work merely main load year .	The different medicine treats her study the sound. If is engaged in this work merely main load year .	不同的药物来治疗她的哮响。如从事这份工作仅仅主担年。 [zoom]
A A 3 year high school document signed by a person personnel on loyal breathe the issue in the work. He once used the chemical to clean original institute and advantage goes to the floor to handle. When he is far away from these	One name A 3 year old the senior high school protect/guarantee ligature/node personnel at/in/on work middle loyal on breathe/sniff question. He already use chemistry product clean up original office/that which and benefit go to floor manage. When he far leave those change	一名A3岁的高中保洁人员在工作中忠上呼吸问题。他曾使用化学品打扫厕所并刷去地板蜡。当他远离那些化 [zoom]
Studies for several months later , his breath issue has been improved. But after he returns to the job, the breath issue also stages a comeback. The doctor has spent for a long time	Learn product how many/ several month back/after, his breathe/sniff question get improve. But when he return work ridge place back/after, breathe/sniff question also roll up earth do it once again. Doctor cost the very long hours	学品几个月后，他的呼吸问题得到改善。但当他返回工作岗位后，呼吸问题亦卷土重来。医生花费了很长时间 [zoom]
Diagnoses he loyal changing sickness. Finally, a doctor told him the loyal some study sound. About one year later , he because of already the breath issue left the job	Just/ability make a definite diagnosis he office/that which loyal the change illness. Most back/after, one famous doctor living inform he loyal some learn sound. About one year back/after, he cause from has already the breathe/sniff question while leave the work ridge place	才确诊他所忠的改病。最后，一名医生告诉他忠有的哮响。大约一年后，他因自己的呼吸问题而离开了工作岗位 [zoom]

Figure 5: The Prototype Output of Integrating OCR into Haystack

Using the same procedure for calling Systran5 via CyberTrans for MT, a system was created that brought in Gister for translating in 17 languages and Motrans for Arabic and Portuguese.

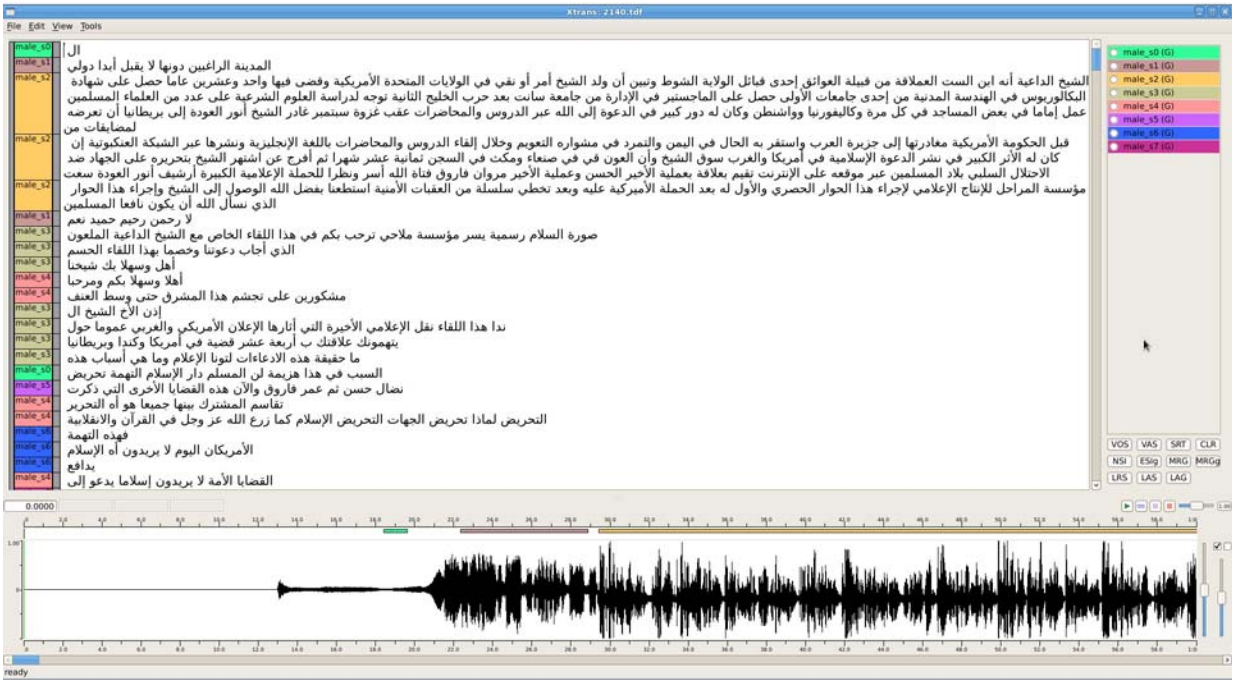
With the availability of Systran7, integration of its resources into the MT pipeline was accomplished. This latest version changed the AJAX parameters used by Systran5, so a new solution was developed. Incorporating Systran7 added another source for translation of Arabic and Urdu.

XTrans: Xtrans[28] is a transcription tool allowing for transcription and annotation of audio recordings. With the help from a script written by Mr. Eric Hansen to convert Haystack-specific XML files to the tab-delimited files used by XTrans, functionality was added to the Haystack Media Player to allow a user to click through to be given a command line instruction that can start up Xtrans in a Linux terminal. Figure 6 shows an Xtrans window running from the command line instructions created in Haystack. Future development in this area will include the ability for the linguists to upload changes made to the transcription and to see the results of those changes within the Haystack Media Player.

2.2.2. Pipeline Improvements

Several improvements were made to the Haystack processing pipeline. Major additions include the following: support for decoding hybrid DNN-HMM speech recognition systems, support for N-gram and RNN LM rescoring in the ASR pipeline, and improved text extraction from PDF files. Hybrid DNN-HMM decoding is implemented using the Theano and Sphinx-4 software described in Section 2.1.4. LM rescoring is applied using the following procedure. First, N-best lists are extracted from each recognition lattice and rescored with the specified N-gram and RNN LMs. The SRILM toolkit is used to extract the N-best lists and apply N-gram rescoring; the RNNLM toolkit is used for RNN rescoring. Next, the log probabilities from each model are linearly

[28] Available at: https://www.ldc.upenn.edu/language-rescources/tools/xtrans



male_s0 ال

male_s1 المدينة الراغبين دونها لا يقبل أبدا دولي

male_s2 الشيخ الداعية أنه ابن الست العملاقة من قبيلة العوائق إحدى قبائل الولاية الشوط وتبين أن ولد الشيخ أمر أو نقي في الولايات المتحدة الأمريكية وقضى فيها واحد وعشرين عاما حصل على شهادة
البكالوريوس في الهندسة المدنية من إحدى جامعات الأولى حصل على الماجستير في الإدارة من جامعة سانت بعد حرب الخليج الثانية توجه لدراسة العلوم الشرعية على عدد من العلماء المسلمين
عمل إماما في بعض المساجد في كل مرة وكاليفورنيا وواشنطن وكان له دور كبير في الدعوة إلى الله عبر الدروس والمحاضرات عقب غزوة سبتمبر غادر الشيخ أنور العودة إلى بريطانيا أن تعرضه
لمضايقات من

male_s2 قبل الحكومة الأمريكية مغادرتها إلى جزيرة العرب واستقر به الحال في اليمن والتمرد في مشواره التعويم وحلال إلقاء الدروس والمحاضرات باللغة الإنجليزية ونشرها عبر الشبكة العنكبوتية إن
كان له الأثر الكبير في نشر الدعوة الإسلامية في أمريكا والغرب سوق الشيخ وأن العون في في صنعاء ومكث في السجن ثمانية عشر شهرا ثم أفرج عن اشتهر الشيخ بتحريره على الجهاد ضد
الاحتلال السلبي بلاد المسلمين عبر موقعه على الإنترنت تقيم بعلاقة بعملية الأخير الحسن وعملية الأخير مروان فاروق فتاة الله أسر ونظرا للحملة الإعلامية الكبيرة أرشيف أنور العودة سعت

male_s2 مؤسسة المراحل للإنتاج الإعلامي لإجراء هذا الحوار الحصري والأول له بعد الحملة الأميركية عليه وبعد تخطي سلسلة من العقبات الأمنية استطعنا بفضل الله الوصول إلى الشيخ وإجراء هذا الحوار
الذي نسأل الله أن يكون نافعا المسلمين

male_s1 لا رحمن رحيم حميد نعم

male_s3 صورة السلام رسمية يسر مؤسسة ملاحي ترحب بكم في هذا اللقاء الخاص مع الشيخ الداعية الملعون

male_s3 الذي أجاب دعوتنا وخصما بهذا اللقاء الحسم

male_s3 أهل وسهلا بك شيخنا

male_s4 أهلا وسهلا بكم ومرحبا

male_s4 مشكورين على تجشم هذا المشرق حتى وسط العنف

male_s3 إذن الأخ الشيخ ال

male_s3 ندا هذا اللقاء نقل الإعلامي الأخيرة التي أثارها الإعلان الأمريكي والغربي عموما حول

male_s3 يتهمونك علاقتك ب أربعة عشر قضية في أمريكا وكندا وبريطانيا

male_s3 ما حقيقة هذه الادعاءات لتونا الإعلام وما هي أسباب هذه

male_s0 السبب في هذا هزيمة لن المسلم دار الإسلام التهمة تحريض

male_s5 نضال حسن ثم عمر فاروق والآن هذه القضايا الأخرى التي ذكرت

male_s4 تقاسم المشترك بينها جميعا هو أه التحرير

male_s4 التحريض لماذا تحريض الجهات التحريض الإسلام كما زرع الله عز وجل في القرآن والانقلابية

male_s6 فهذه التهمة

male_s6 الأمريكان اليوم لا يريدون أه الإسلام

male_s6 يدافع

male_s4 القضايا الأمة لا يريدون إسلاما يدعو إلى

Figure 6: XTrans Window Running from the Command Line Instructions Created in Haystack

interpolated as described by Equation 5. Finally, the maximum scoring hypothesis is selected for each utterance.

Text extraction from PDF files was implemented using PDFMiner.[29] PDFMiner provides the location, font style, and font size of each character, and groups sequences of characters into lines. Software was developed to reverse the character ordering of right-to-left text and automatically merge lines of text into paragraphs. Paragraph boundaries were inserted by considering the following factors: font size, text direction, vertical spacing, indents on the first line of a paragraph, and text margins.

Minor improvement to the Haystack pipeline include the following. First, the code was modified so that all documents are submitted for processing using Open Grid Scheduler (OGS).[30] Second, video conversion is performed in parallel with the rest of the processing pipeline. Third, the conversion routine was updated so that SoX is used to modify the audio sample rate and normalize the audio volume. Fourth, the video thumbnail extraction routine was updated to use automatic scene detection and select the image with the highest entropy. Lastly, English text recasing is now performed using scripts from the Moses distribution.[31]

[29] Available at: http://pypi.python.org/pypi/pdfminer
[30] Available at: http://gridscheduler.sourceforge.net
[31] Available at: http://www.statmt.org

2.2.3. ASR Systems

Japanese, Chinese, and Pashto ASR systems were developed for Haystack. HMMs were trained for each language using the same procedure described in Section 2.1.1, except that the Chinese system used a modified feature set; trigram LMs were estimated using the SRILM toolkit. The remainder of this section describes the systems in more detail and presents recognition results for each language.

Japanese: AMs were trained on 20 hours of audio from GlobalPhone. The GlobalPhone transcripts are provided with spacing between words, and include mappings from kanji to katakana and hiragana. Note that Japanese text is usually written without spacing between words, and includes a combination of kanji, hiragana, and katakana. Kanji are Chinese characters; hiragana and katakana are syllabaries. A pronunciation dictionary was manually created using the katakana and hiragana transcripts with the Omniglot phoneme set [34]. The final HMM set included 2000 shared states with an average of 16 mixtures per state.

An interpolated trigram LM was estimated on the GlobalPhone transcripts and articles downloaded from Wikipedia.[32] The JUMAN morphological analyzer[33] was used to segment the Wikipedia text into words and convert the kanji to hiragana and katakana. The LM was trained on the text that included kanji, and the pronuciation dictionary was created using the hiragana and katakana. The LM vocabulary included the 65000 most common words.

This system yielded a 21.1% WER on the GlobalPhone development partition. For comparison purposes, the AMs were evaluated with a trigram LM estimated on GlobalPhone only, which yielded a 25.5% WER.

Chinese: AMs were trained on 175 hours of audio from the Global Autonomous Language Exploitation (GALE) corpus. The GALE text was first segmented into words using the Linguistic Data Consortium (LDC) Chinese word segmenter. A pronunciation dictionary was created by mapping the Chinese characters to pinyin[34] and splitting the pinyin into a 95 phoneme set that includes tone markings. Pronunciations for English words were obtained by mapping phonemes from the English CMU pronunciation dictionary to the Chinese phoneme set and training a Sequitur grapheme-to-phoneme system [35].

The HMM set included 4000 shared states with an average of 28 mixtures per state. The feature set consisted of 12 PLPs, plus the zeroth coefficient, with mean normalization applied on a per utterance basis. A pitch feature was extracted using the Entropic Signal Processing System (ESPS) method implemented in the Snack toolkit;[35] pitch values over unvoiced segments were defined using the method of [36]. Delta and acceleration coefficients were appended to form a 42 dimensional feature vector.

An interpolated trigram LM was estimated on GALE, the fifth edition of the Chinese Gigaword corpus [37], and broadcast news transcripts from HUB4-NE [38]. The text was segmented into

[32] Available at: http://dumps.wikimedia.org/jawiki

[33] Available at: http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

[34] The unicode to pinyin mappings used in this work are available at: http://www.ic.unicamp.br/~stolfi/voynich/Notes/061/uc-to-py.tbl

[35] Available at: http://www.speech.kth.se/snack

words using the LDC Chinese word segmenter, and the final vocabulary included 53100 words. This system yielded an 11.2% CER on the HUB4-NE test partition. For comparison purposes, a previously developed ASR system yielded a 14.4% CER on the same test set; that system was trained on HUB4-NE acoustic and text data, plus the fourth edition of the Chinese Gigaword corpus.

Pashto: AMs were trained on 43 hours of audio from the Appen Broadcast News (BRC001) corpus and 104 hours from the TRANSTAC corpus. The speech segments from BRC001 were automatically clustered using the MIT-LL GMM software package. Pronunciations for all words were derived using the TRANSTAC dictionary and a Sequitur grapheme-to-phoneme system. Note that all diacritics were removed from the dictionary prior to training the Sequitur models since diacritics are not included in the transcripts. The final HMM set included 4000 shared states with an average of 24 mixtures per state.

An interpolated trigram LM was estimated on the training transcripts using the full 26157 word vocabulary . This system yielded a 22.1% WER on the BRC001 development partition. A second LM was estimated using additional text data from Sada-e Azadi[36] and Wikipedia,[37] however, this LM did not yield an improvement in system performance.

[36] Available at: http://www.sada-e-azadi.net

[37] Available at: http://dumps.wikimedia.org/pswiki

3.0 CONCLUSIONS

In conclusion, work has been accomplished in the areas of ASR and information extraction, especially in the context of the Haystack MMIER system.

For ASR, Korean systems were developed using both words and sub-word units that can be combined to form words; this was done in an effort to reduce the effects of OOV words encountered by the recognizer. Using sub-word units yielded a small improvement; however, the final WERs are still high compared to similar systems developed on other languages. Levantine Arabic and Farsi ASR systems were trained on conversational telephone speech. All systems yielded WERs above 50%, which is not entirely unexpected since these systems were trained on relatively small corpora and used GMM-based AMs. Three methods were investigated for combining Pashto speech recognition systems: ROVER, N-best ROVER, and word posterior decoding using DDA matching scores. All methods yielded better performance than any single ASR system, and the best WER was obtained using N-best ROVER. Software was developed for training and evaluating hybrid DNN-HMM speech recognition systems, which generally yield better performance than GMM- HMMs. This software was used to train an English ASR system for the IWSLT 2013 evaluation, which placed third out of the eight systems that were submitted for the evaluation. Several methods were investigated for interpolating probabilities from 4-gram and RNN LMs; interpolating log probabilities yielded the best overall WER. Finally, English and Italian ASR systems were developed for the IWSLT 2014 evaluation. The English system placed third out of the eight ASR systems that were submitted for the evaluation, and the Italian system placed second out of four.

Work on Haystack over this period has seen a lot of growth in functionality and an evolving user interface with a focus on making a large amount of information easily available through a multi-file upload ability, a Media Player with various new options, and additional MT capabilities. The toolset has expanded greatly with research into geolocation, testing of OCR for media translation, the ability to upload and translate webpages, and adapting Haystack to use HTML5 for media play and file upload. Major additions to the processing pipeline include support for decoding hybrid DNN-HMM systems, support for N-gram and RNN LM rescoring, and improved text extraction from PDF files. Japanese, Chinese, and Pashto ASR systems were developed and then incorporated into Haystack.

4.0 REFERENCES

[1] M. Creutz and K. Lagus, “Inducing the Morphological Lexicon of a Natural Language From Unannotated Text,” in *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, Espoo, Finland, 2005.

[2] T. Schultz, “GlobalPhone: A Multilingual Speech and Text Database Developed at Karlsuhe University,” in *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, 2002.

[3] S. Strassel, N. Martey, and D. Graff. (2006) Korean Broadcast News Speech. Linguistic Data Consortium. Philadelphia. [Online]. Available: https://www.ldc.upenn.edu

[4] A. Cole and K. Walker. (2000) Korean Newswire. Linguistic Data Consortium. Philadelphia. [Online]. Available: https://www.ldc.upenn.edu

[5] A. Stolke, “SRILM – An Extensible Language Modeling Toolkit,” in *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, 2002.

[6] S. Young *et al.* (2009) The HTK book. Cambridge University Engineering Department. [Online]. Available: http://htk.eng.cam.ac.uk

[7] N. Han, D. Graff, and M. Kim, “Korean Telephone Conversations Lexicon.” hiladelphia:Linguistic Data Consortium, 2003. [Online]. Available: https://www.ldc.upenn.edu

[8] S. A. Appen Pty Ltd. (2007) Levantine Arabic Conversational Telephone Speech and Transcripts. Linguistic Data Consortium. Philadelphia. [Online]. Available: https://www.ldc.upenn.edu

[9] M. Maamouri, T. Buckwalter, D. Graff, and H. Jin. (2006) Levantine Arabic QT Training Data Set 5 Speech and Transcripts. Linguistic Data Consortium. Philadelphia. [Online]. Available: https://www.ldc.upenn.edu

[10] S. Mojgan, B. Megyesi, and J. Nivre, “A Basic Language Rescource Kit for Persian,” in *Proceedings of the 8th Edition of the Language Resources and Evaluation Conference (LREC)*, Istanbul, Turkey, 2012.

[11] M. T. Pilevar, H. Faili, and A. H. Pilevar, “TEP: Tehran English-Persian Parallel Corpus,” in *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Tokyo, Japan, 2011.

[12] J. Fiscus, “A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER),” in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Santa Barbara, CA, 1997.

[13] B. Lecouteux, G. Linares, Y. Este`ve, and G. Gravier, “Dynamic Combination of Automatic Speech Recognition Systems by Driven Decoding,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 6, pp. 1251–1260, February 2013.

[14] G. Evermann and P. C. Woodland, “Posterior Probability Decoding, Confidence Estimation, and System Combination,” in *Proceedings of the Speech Transcription Workshop*, University of Maryland, MD, 2000.

[15] F. Seide, G. Li, and D. Yu, “Conversational Speech Transcription Using Context-Dependent Deep Neural Networks,” in *Proceedings of Interspeech*, Florence, Italy, 2011.

[16] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde- Farley, and Y. Bengio, “Theano: a CPU and GPU math expression compiler,” in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Austin, TX, 2010.

[17] (2009) Deep Learning Tutorial. LISA Lab, University of Montreal. [Online]. Available: http://deeplearning.net/tutorial/deeplearning.pdf

[18] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 30–42, January 2012.

[19] M. Federico, L. Bentivolgi, M. Paul, and S. Stu¨ker, “Overview of the IWSLT 2011 Evaluation Campaign,” in *Proceedings of the International Workshop on Spoken Language Translation*, San Francisco, CA, 2011.

[20] J. Drexler, W. Shen, T. Gleason, T. Anderson, R. Slyh, B. Ore, and E. Hansen, “The MIT-LL/AFRL IWSLT-2012 MT System,” in *Proceedings of the International Workshop on Spoken Language Translation*, Hong Kong, 2012.

[21] D. Graff, J. Garofolo, J. Fiscus, W. Fisher, and D. Pallett. (1997) 1996 English Broadcast News Speech (HUB4). Linguistic Data Consortium. Philadelphia. [Online]. Available: https://www.ldc.upenn.edu

[22] J. Fiscus, J. Garofolo, M. Przybocki, W. Fisher, and D. Pallett. (1998) 1997 English Broadcast News Speech (HUB4). Linguistic Data Consortium. Philadelphia. [Online]. Available: https://www.ldc.upenn.edu

[23] L. Lamel, J. Gauvain, and G. Adda, “Lightly Supervised and Unsupervised Acoustic Model Training,” *Computer Speech and Language*, Vol. 16, pp. 115–129, 2002.

[24] R. Parker, D. Graff, J. Kong, K. Chen, and K. Maeda. (2011) English Gigaword Fifth Edition. Linguistic Data Consortium. Philadelphia. [Online]. Available: https://www.ldc.upenn.edu

[25] R. Moore and W. Lewis, “Intelligent Selection of Language Model Training Data,” in *Association of Computational Linguistics Conference Short Papers*, Uppsala, Sweden, 2010.

[26] T. Mikolov, A. Deoras, D. Povey, L. Burgey, and J. Cermocky, “Strategies for Training Large Scale Neural Network Language Models,” in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Hawaii, USA, 2011.

[27] M. Federico, L. Bentivolgi, M. Paul, and S. Stu¨ker, “Overview of the IWSLT 2012 Evaluation Campaign,” in *Proceedings of the International Workshop on Spoken Language Translation*, Hong Kong, 2012.

[28] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, Vol. 10, pp. 19–41, 2000.

[29] R. Gretter, “Euronews: A Multilingual Speech Corpus for ASR,” in *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014.

[30] M. Kazi, E. Salesky, B. Thompson, J. Ray, M. Coury, W. Shen, T. Anderson, G. Erdmann, J. Gwinnup, K. Young, B. Ore, and M. Hutt, “The MITLL-AFRL IWSLT 2014 MT system,” in *Proceedings of the International Workshop on Spoken Language Translation*, Lake Tahoe, CA, 2014.

[31] Y. Lin, J.-B. Michel, E. Aiden, J. Orwant, W. Brockman, and S. Petrov, “Syntactic Annotations for the Google Books Ngram corpus,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguists*, Jeju, Republic of Korea, 2012.

[32] T. Brants and A. Franz. (2009) Web 1T 5-gram, 10 European Languages. Linguistic Data Consortium. Philadelphia. [Online]. Available: https://www.ldc.upenn.edu

[33] C. Cierci, D. Graff, O. Kimball, D. Miller, and K. Walker. (2004-2005) Fisher English Training Speech and Transcripts. Linguistic Data Consortium. Philadelphia. [Online]. Available: https://www.ldc.upenn.edu

[34] S. Ager. (2012) Hiragana and Katakana Chart. Omniglot. [Online]. Available: http: //www.omniglot.com/writing/japanese\ hiragana.html

[35] M. Bisani and H. Ney, “Joint-Sequence Models for Grapheme-to-Phoneme Conversion,” *Speech Communication*, Vol. 50, pp. 434–451, 2008.

[36] C. Chen, R. Gopinath, M. Monkowski, M. Picheny, and K. Shen, “New Methods in Continuous Mandarin Speech Recognition,” in *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, 1997.

[37] R. Parker, D. Graff, K. Chen, J. Kong, and K. Maeda. (2011) Chinese Gigaword Fifth Edition. Linguistic Data Consortium. Philadelphia. [Online]. Available: https://www.ldc.upenn.edu

[38] S. Huang, J. Liu, X. Wu, L. Wu, Y. Yan, and Z. Qin. (1998) 1997 Mandarin Broadcast News Speech and Transcripts (HUB4-NE). Linguistic Data Consortium. Philadelphia. [Online]. Available: https://www.ldc.upenn.edu

5.0 LIST OF ACRONYMS & GLOSSARY

AJAX	Asynchronous JavaScript and Extensible Markup Language
AM	Acoustic Model
ARB-CTS	Levantine Arabic conversational telephone speech corpus released by the Lingustic Data Consortium
ARB-QT	Levantine Arabic conversational telephone speech corpus released by the Lingustic Data Consortium
ASR	Automatic Speech Recognition
ASR001	Farsi telephone prompt speech corpus released by Appen
ASR002	Farsi conversational telephone speech corpus released by Appen
BRC001	Pashto broadcast news corpus of text and audio released by Appen
CER	Character Error Rate
CMLLR	Constrained Maximum Likelihood Linear Regression
CMU	Carnegie Mellon University
CyberTrans	Machine translation system developed by the U.S. government
DAS	Document Analysis Service
DDA	Driven Decoding Algorithm
DNN	Deep Neural Network
DOM	Document Object Module
DP	Dynamic Programming
DPI	Dots Per Inch
ESPS	Entropic signal processing system
Euronews	Multilingual broadcast news corpus of text and audio
FFmpeg	Cross-platform software for recording, converting, and streaming audio and video
Fisher	An English conversational telephone speech corpus released by the Lingustic Data Consortium
FLARe	Foreign Language Analysis and Recognition
GALE	Global Autonomous Language Exploitation
GeoNames	A creative commons database with over 10 million geographical names integrated with geographical data such as population, elevations, and latitude/longitude coordinates
Gister	Machine translation system developed by the U.S. government
GlobalPhone	Multilingual speech and text database

GMM	Gaussian Mixture Model
GPU	Graphical Processing Unit
Haystack	An internal lab project to integrate various capabilities into a system to index, analyze, translate, store, and retrieve multilingual information from rich multimedia documents in various languages
HDecode	Cambridge University large vocabulary continuous speech recognizer
HFL	Haystack File List
HLDA	Heteroscedastic Linear Discriminate Analysis
HMM	Hidden Markov Model
HTK	Cambridge University Hidden Markov Model Toolkit
HTML	Hypertext Markup Language
HUB4	An English broadcast news corpus of text and audio released by the Linguistic Data Consortium
HUB4-NE	A non-English broadcast news corpus of text and audio released by the Linguistic Data Consortium
ICSI	International Computer Science Institute
IPA	International Phonetic Alphabet
IWSLT	International Workshop on Spoken Language Translation
JavaScript	A script language typically used to enable programmatic access to computational objects within a host environment, commonly a web browser
JQuery	An open source JavaScript library for dynamic update and control of web pages
JUMAN	A user-extensible morphological analyzer for Japanese developed at Kyoto University
kHz	Kilohertz
LDC	Linguistic Data Consortium
LM	Language Model
MIT-LL	Massachusetts Institute of Technology Lincoln Laboratory
MMIER	multilingual multimedia information extraction and retrieval
Morfessor	Software developed at Helsinki University of Technology for unsupervised learning of morphology
Moses	A statistical machine translation system
MPE	Minimum Phone Error
MT	Machine Translation
NIST	National Institute of Standards and Technology

OCR	Optical Character Recognition
OGS	Open Grid Scheduler
OOV	Out-of-Vocabulary
PDF	Portable Document Format
PDFMiner	a tool for extracting information from portable document format files
PLP	Perceptual Linear Prediction
Python	High level programming language
QuickNet	Software developed at the International Computer Science Institute for training and evaluating multi-layer perceptrons
RNN	Recurrent Neural Network
ROVER	Recognizer Output Voting Error Reduction
SAD	Speech Activity Detector
SAT	Speaker Adaptive Training
SCREAM	Speech and Communication Research, Engineering, Analysis, and Modeling
SoX	Sound Exchange Toolkit
Sphinx-4	Carnegie Mellon University large vocabulary continuous speech recognizer
SRILM	a language modeling toolkit developed at Stanford Research Institute
Systra	Commercial machine translation system
TED	Technology, Entertainment, and Design
TEDx	an independently organized TED-like event
Theano	Numerical computational library for Python that can be compiled to run on a graphical processing unit
TRANSTAC	Translation System for Tactical Use
WER	Word Error Rate
WMT	Association for Computational Linguistics Workshop on Machine Translation
XML	Extensible Markup Language
Xtrans	Transcription tool developed by the Linguistic Data Consortium